AD/A-007 247

TIME AND FREQUENCY RESOLUTION IN SPEECH ANALYSIS AND SYNTHESIS

Aubrey M. Bush

Georgia Institute of Technology

Prepared for:

Army Research Office-Durham

10 January 1975

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER E21-611-74-BU-2	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER AD/A-007247
4. TITLE (and Subtitle) Time and Frequency Resolution in Speech Analysis and Synthesis		5. TYPE OF REPORT & PERIOD COVERED Final
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Aubrey M. Bush		8. CONTRACT OR GRANT NUMBER(s) DA-ARO-D-31-124-71-G126
9. PERFORMING ORGANIZATION NAME AND ADDRESS Georgia Institute of Technology School of Electrical Engineering		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Office Box CM Duke Station Durham, NC 27706		12. REPORT DATE January 10, 1975
		13. NUMBER OF PAGES 33
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Same		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Unlimited, Open Publication

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

Digital Signal Processing, Speech Compression, Vocoders, Digitization, Speech Quality, Digital Speech Transmission

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

A problem basic to the development of all digital telecommunication systems is the efficient digital encoding of speech signals for transmission. Many speech digitization algorithms have been proposed. This study, using as a research vehicle the homomorphic vocoder algorithm, is directed toward determining the time resolution and the frequency resolution required to faithfully reproduce a speech signal. The results reported here are fundamental in that they are applicable to any speech digitization algorithm.

PTO

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT (Cont'd)

The resolutions required for typical speaker situations are summarized.

It was demonstrated that the dominant feature of speech analysis-synthesis digitization algorithms is the pitch extraction; pitch errors often lead to gross distortion of the reconstructed speech.

TIME AND FREQUENCY RESOLUTION IN SPEECH ANALYSIS AND SYNTHESIS

FINAL REPORT

JANUARY 10, 1975

U. S. ARMY RESEARCH OFFICE - DURHAM

GRANT NO. DA-ARO-D-31-124-71-G126

GEORGIA INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING
E21-611-74-BU-2

THE FINDINGS IN THIS REPORT ARE NOT TO BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, UNLESS SO DESIGNATED BY OTHER AUTHORIZED DOCUMENTS.

CONTENTS

STATEMENT OF THE PROBLEM

Speech Production[1]

The sounds of human speech are produced when the vocal tract (an acoustic cavity) is excited by a flow of air from the lungs. Voiced speech results when air is forced through the glottis (the opening between the vocal cords) while the vocal cords are held under tension. The glottis oscillates causing a quasi-periodic flow of air to excite the vocal tract. The glottal signal is a pulse-train time-function, rich in harmonics. The fundamental frequency of vocal cord oscillation is called the voice pitch.

The excitation signal from the glottis passes through the vocal tract, which includes the throat, mouth, and nasal cavity. The message the talker wants to convey is imposed on the excitation signal by the changes in position of the tongue, lips, and other moving parts of the tract. These moving parts are called articulators, and their activity in creating the spoken language is called articulation. During articulation the vocal cavity assumes different positions causing resonances in the tract which alter the spectrum of the excitation signal, imposing on the spectrum peaks which are called formants.

Unvoiced speech is produced by a turbulent flow of air past a constriction in the vocal tract, or by a release of pressure at some point of closure in the tract. Unvoiced excitation is an acoustic noise source. The spectrum of an unvoiced speech sound is influenced mainly by that portion of the vocal tract forward of the constriction. Pressure released at a closure causes an initial burst, followed by turbulent flow noise.

The phoneme is the smallest unit of speech that distinguishes one utterance from another. General American English has about 42 phonemes [1]. We may think of these phonemes as a code uniquely related to the articulatory gestures of the language.

The vowel sounds of speech are produced by voiced excitation of the vocal tract (e.g., the "ah" in father). In normal articulation the tract is held in a relatively stable position during most of the sound. Vowels usually have a "duration" of 60 ms or longer.

The fricative consonant phonemes are produced by incoherent noise excitation of the tract caused by turbulent air flow at a constriction (e.g., the "s" in see). The vocal cord source may operate in conjunction with the noise source to produce a voiced fricative (e.g., the "z" in zoo).

Stop consonants are produced by the abrupt release of pressure at a place of closure in the tract (e.g., the "t" in to). The articulatory movements which generate stops are more rapid than for other sounds. Stops may be voiced or unvoiced.

The remaining consonants are classified as nasals, glides, semivowels, dipthongs, and affricates.

A key observation about the phonemes of speech is that some sounds (e.g., the stop consonants) are produced by a rapid motion of the articulators, while others (e.g., the vowels) are produced by a relatively stable vocal tract configuration. The striking difference between the character of "short" sounds are "long" sounds suggests that all sounds should not be processed in the same manner.

Spectrum Analysis

The traditional tool for spectrum analysis of signals and linear systems is the Fourier transform pair:

$$S(f) = \int_{-\infty}^{\infty} s(f)\epsilon^{-j2\pi ft}\,dt \qquad s(t) = \int_{-\infty}^{\infty} S(f)\epsilon^{+j2\pi ft}\,df \tag{1}$$

In analyzing a speech signal, however, the future is not available, and only the very recent past is of interest. So we adopt the short-time spectrum $S(t,f)$ [1]:

$$S(t,f) = \int_{t-D}^{t} s(\tau)w(t-\tau)\epsilon^{-j2\pi f\tau}\,d\tau\ , \tag{2}$$

where D is the duration of the window function $w(t)$. The short-time spectrum is the Fourier transform of the recent past of the time function $s(\tau)$ weighted by the window function $w(t-\tau)$. Thus, $S(t,f)$ describes the distribution of energy in frequency, as it changes with time. The generation and coding of this short-time spectrum are the central features of most speech data rate reduction systems.

The short-time spectrum may be displayed with a sound spectogram--a representation of the time-frequency-intensity coordinates of $|S(t,f)|$ [1,5]. The display is generated by playing a recorded passage of speech (typically 2.4 seconds) through a narrow band-pass filter and envelope-detector. The contour $|S(t,f_i)|$ is "burned" on Teledeltos paper for successive filter center-frequencies f_i, with relative darkness displaying intensity on a logarithmic scale. Two analyzing filter bandwidths are commonly available--45Hz and 300Hz. The choice of filter

bandwidth corresponds (roughly) to the choice of the duration of the window function in (2).

In the narrow-band mode, the frequency resolution is sufficient to display the voice pitch and its harmonics, but the time resolution is relatively poor. In the wide-band mode, the time resolution is sufficient to display individual glottal pulses, but the frequency resolution is relatively poor. The formant structure of the speech signal may be observed in voiced portions of the spectograms.

Notice the effect of the window function on the short-time spectrum. The scaling property of the Fourier transform

$$s(t) \leftrightarrow S(f) \Rightarrow s(at) \underset{t \quad f}{\leftrightarrow} \frac{1}{|a|} S(\frac{f}{a}) \tag{3}$$

shows that as the effective "duration" of a window function $w(t)$ is made shorter, the "bandwidth" of its spectrum $W(f)$ is broadened, and vice versa. Since multiplication of the signal $s(\tau)$ by the window $w(t-\tau)$ is equivalent to convolving their spectra it is clear that the spectrum of a sinusoid viewed through a time window is broadened as the duration of the window is decreased. The choice of a window function involves a compromise between the time "resolution" and frequency "resolution" that may be achieved in the short-time spectrum [5].

Digital spectrum analysis is accomplished with the discrete Fourier transform (DFT) pair:

$$S(kF) = \sum_{n=0}^{N-1} s(nT)\epsilon^{-j2\pi nk/N} \qquad s(nT) = \frac{1}{N}\sum_{k=0}^{N-1} S(kF)\epsilon^{+j2\pi nk/N} \tag{4}$$

where T is the sampling interval of the time function s(t), N is the number of samples to be transformed, $F = \frac{1}{NT}$ is the sampling interval of the spectrum [8].

To obtain the discrete short-time Fourier transform, we introduce the window function into (4):[5]

$$S_r(kF) = \sum_{n=0}^{N-1} w(nT)\ s(nT + rMT)\epsilon^{-j2\pi nk/N} . \tag{5}$$

The index r corresponds to the time variable in (2). The short time spectrum is evaluated at times t = rMT, for r = 0,1,2..... The window is propagated along the time function S(t) in steps of MT seconds. The samples of (5) represent the samples of (2) to within a phase constant, i.e.

$$|S_r(kF)| = |S(rMT + D,\ kF)| . \tag{6}$$

The discrete short-time Fourier transform may be efficiently computed digitally with the Fast Fourier Transform Algorithm.[8][9]

The short time spectrum can also be plotted by digital means either for display on a CRT or a Calcomp or similar plotter, using "3-D" plotting routines. In this case, a wider choice of parameters is available.

Linear Prediction

Techniques for speech analysis which separate the vocal tract response function and the excitation signal using time domain techniques rather than concentrating initially on a short time spectrum are generally lumped into a class of system referred to in the literature as "LPC's" or linear predictive coders.[10] This class of speech analysis operates

as one of several possible forms of least squares prediction algorithms.[11] The vocal tract is assumed to be an all pole filter, excited by an impulse train or white noise. The prediction algorithm is chosen to locate the poles of the vocal tract filter. The vocal tract prediction is then subtracted from the signal to leave a residual containing primarily excitation information. Although algorithms which seek to determine structural details of filters are in general nonlinear [12], utilization of an inverse filter format allows this problem to become a linear optimization algorithm. [9]

The input signal, preparatory to the prediction, is first placed in digital format. Then, depending on the particular prediction chosen, some form of windowing strategy is chosen. The optimization may be either a block algorithm or a point-by-point algorithm. If the block strategy, which is the most common technique, is selected, a window and framing operation must be defined explicitly. The window may be rectangular, Hamming, Hanning [13], etc. The frame interval is usually equal to the window length but may be longer or shorter. There may be an overlap of windows in the framing process. A variety of initialization techniques may be used at each step [14]. If the point-by-point [15] approach is chosen, the algorithm itself will implicitly define a windowing of the signal. In this case, the window is a sliding window, and is generally an exponential window.

Extensive work on the prediction algorithms has been undertaken and reported in the literature.[16][17],[10][11],[15],[18]-[20] Attention has not been directed toward best windowing and framing strategies. The fact that, even though the LPC is a "time domain" approach, a frequency

resolution and time resolution are nevertheless determined by the parameters chosen in the algorithm is not generally appreciated.

Time-Frequency Resolution

The primary concern in this research has been to determine, via subjective listening tests in a controlled laboratory environment, the time and frequency resolution required to faithfully reproduce speech signals.

This was accomplished by processing speech signals using as a research vehicle the homomorphic vocoder described below.

TECHNIQUES, TOOLS, AND PROCEDURES

The Homomorphic Vocoder

The spectrum of a quasi-stationary segment of speech is the product of the excitation spectrum E(f) and the vocal tract system function H(f).

$$s(t) \underset{t \quad f}{\longleftrightarrow} S(f) = E(f)\ H(f)$$

The logarithm of the amplitude spectrum $|S(f)|$ is

$$\ln|S(f)| = \ln|E(f)| + \ln|H(f)| \tag{7}$$

in which the influence of source and vocal tract are additive. Taking the Fourier transform of $\ln|S(f)|$ yields the so-called "cepstrum" $C(\tau)$[4],[8]:

$$C(\tau) = \mathcal{F}\{\ln|S(f)|\} = \int_{-\infty}^{\infty} \ln|S(f)|\,\epsilon^{-j2\pi f\tau}\,df$$

$$= \mathcal{F}\{\ln|E(f)|\} + \{\ln|H(f)|\}\ . \tag{8}$$

During voiced sounds the two components of the cepstrum occupy different regions in the "quefrency" variable τ.

Since $|H(f)|$ is a "smooth" function of frequency, so is $\ln|H(f)|$, and the contribution to the cepstrum is essentially confined to the low-quefrency region on the τ axis. On the other hand, for a voiced sound $|F(f)|$ is essentially periodic in f (with peaks separated by the pitch

frequency f_c) so the contribution to the cepstrum is a spike at $\tau_0 = 1/f_0$, the pitch period.

It is clear that to the extent to which the cepstrum of E(f) and H(f) are disjoint in quefrency τ the vocal tract influence may be isolated by low quefrency filtering, while the excitation may be characterized by a measure of the pitch period τ_0 in the high quefrency region. This describes the strategy of the homomorphic vocoder.

It has been reported that samples of the cepstrum out to $\tau = 3$ ms are adequate for speech signals.[4] Thus, in the digital homomorphic vocoder, the vocal tract spectrum is encoded by the low quefrency cepstrum samples, while voicing information is transmitted as a voiced/unvoiced decision and the pitch frequency f_0 determined from the peak in the cepstrum.

A block diagram of the homomorphic vocoder is shown in Figure 1. Typical waveforms encountered in this vocoder are shown in Figure 2.

The time and frequency resolution of the homomorphic analysis technique are directly controlled. A detailed discussion has been developed by Patisaul.[21] The time resolution is determined by the width of the window in the time domain. The frequency resolution is determined by the length of the window in the cepstral domain. Under suitable conditions, generally appropriate to speech analysis,[21] the time and frequency resolutions are controlled essentially independently.

Subjective Testing

A set of four test sentences taken from the Harvard list of phonetically balanced sentences spoken by a male and by a female speaker

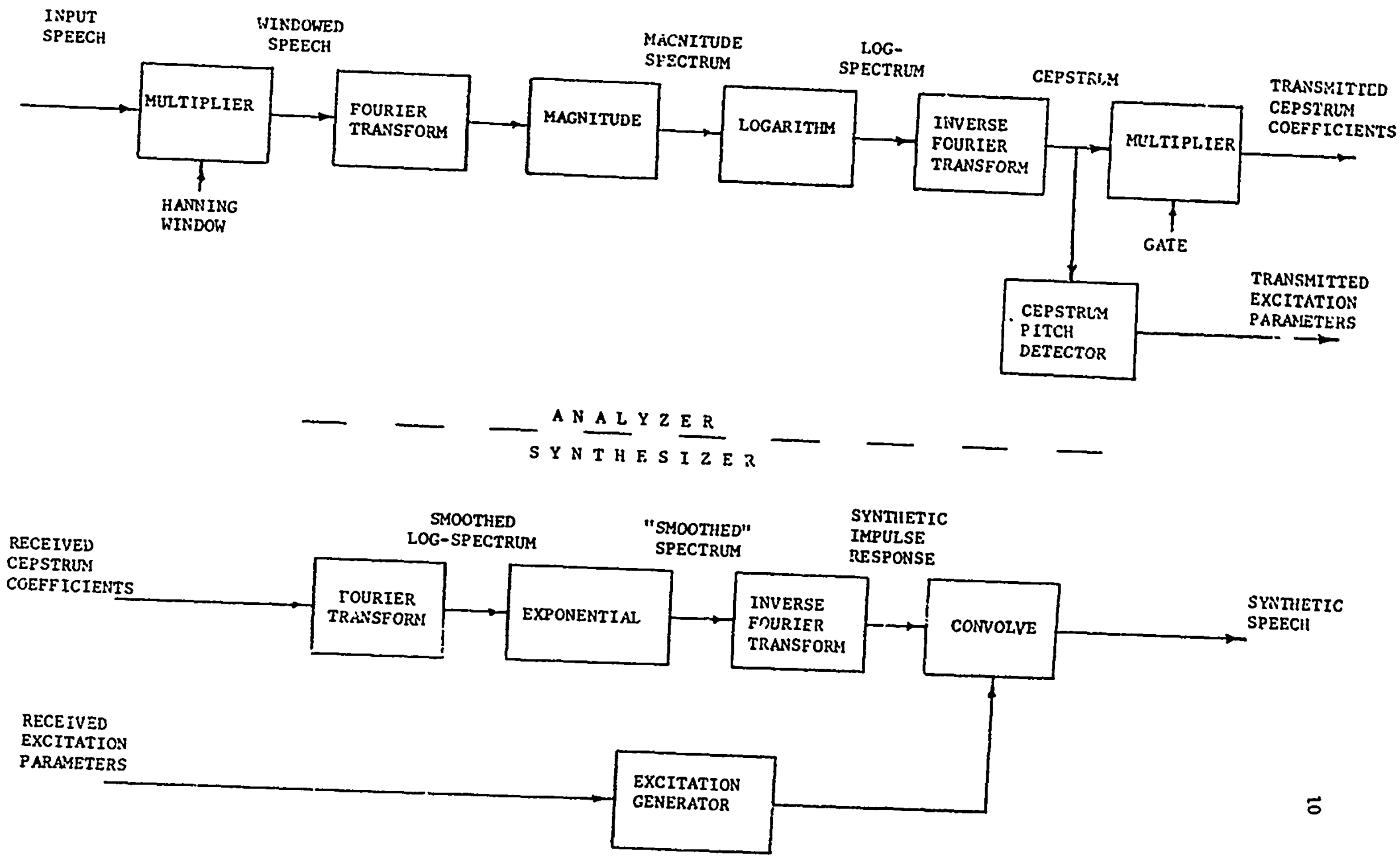

Figure 1 The Homomorphic Vocoder.

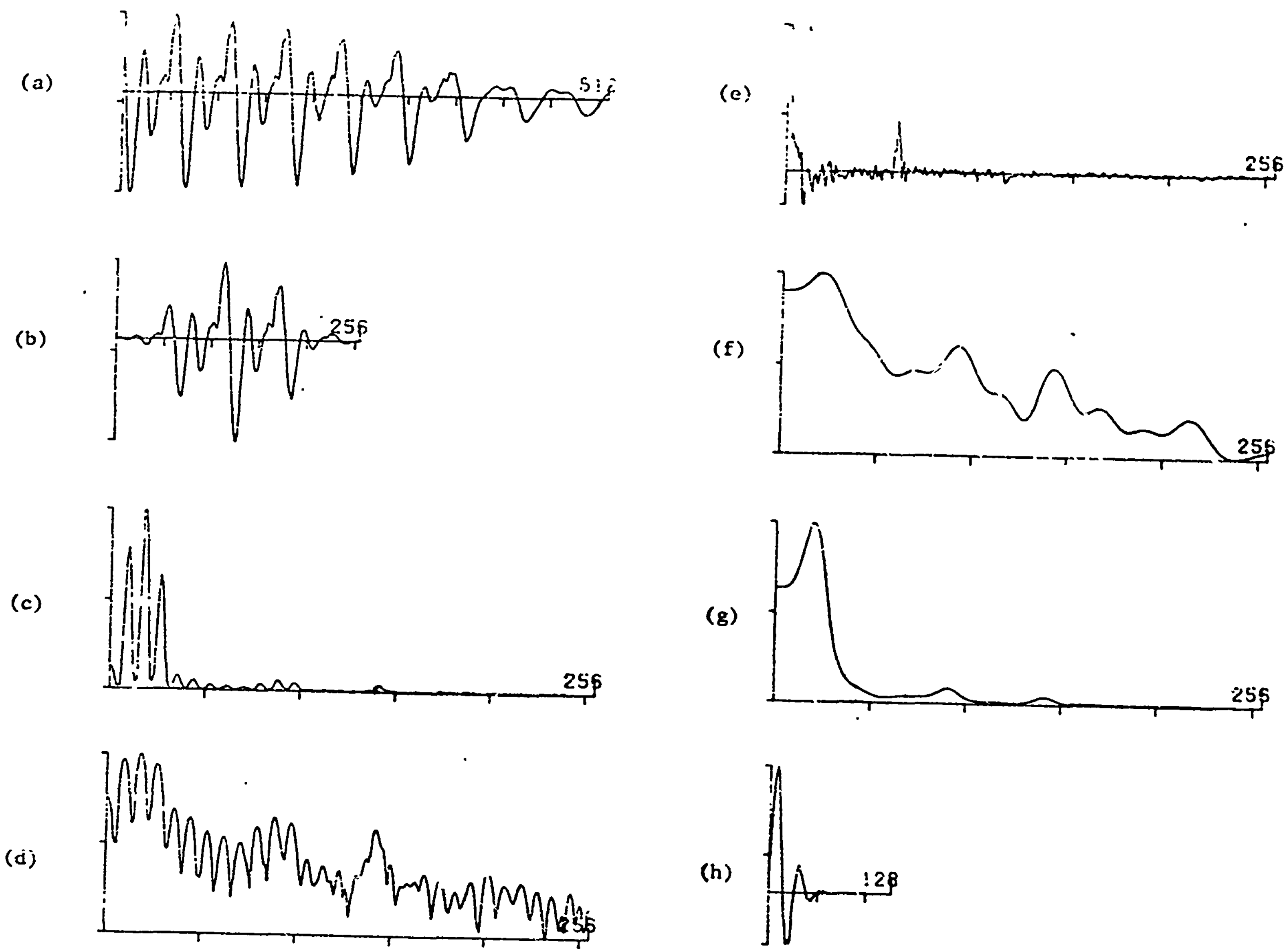

Figure 2 - TYPICAL WAVEFORMS IN THE HOMOMORPHIC VOCODER: (a) The input speech (a 51.2 ms segment of a voiced sound); (b) The windowed speech (25.6 ms window); (c) The magnitude spectrum (scale 0-5KHz); (d) The log-spectrum; (e) The cepstrum (0-25.6 ms); (f) The smoothed log-spectrum; (g) The "smoothed" spectrum; (h) The synthetic impulse response.

was used as source material. These sentences were processed by a homomorphic vocoder utilizing a wide range of windows in both the time and cepstral domains. Fixed windowing strategies utilizing a given combinations of windows throughout were studied as well as adaptive strategies using one combination of windows for voiced segments and another for unvoiced segments. Each combination of windows is referred to as a mode of operation of the vocoder.

A category preference test was chosen for evaluation of the test sentence quality. The results of the category preference judgements was analyzed in great detail to determine significant features.

Each listener in the Category Judgement test is required to rank the test sentences on a scale of 0-8 compared to a "good" and "poor" sample which are presented periodically throughout the test for reference. A mean category judgement (MCJ) can be obtained by averaging across all listeners.

This testing technique is felt to be nearer real world conditions than tests involving direct one-to-one comparisons.

The Simulation System

The system used to implement the homomorphic vocoder and conduct the listening tests has been developed over the past three years.

Initially, simulations were run on a central computer facility, a Univac U1108. Input was accomplished via a Radiation, Inc. A/D unit at the central facility; output was taken off on paper tape and D/A conversion accomplished on a PDP-8 minicomputer located at another site.

Improvement was made by the middle of the first year of the project by substitution of a magnetic tape link for the paper tape link. A Honeywell H-316 was used in place of the PDP-8.

However, input/output constraints were still so severe that no user interaction could be allowed and progress in running simulations was extremely slow.

During the second year of the project a dedicated minicomputer system was secured for our speech research work. Substantial effort was devoted to the development of this system over this period.

The system is uniquely suited to speech research. It is highly interactive and has the memory and peripherals required to efficiently conduct speech research.[22]

All of the results reported below were obtained in the third year of the project using this dedicated minicomputer facility.

Subjective testing was carried out using a professional quality audio system in a room having a controlled environment, connected to the computer room via coaxial cables.

SUMMARY OF RESULTS

Goal

The goal of this research was to study the effects of vocoder time-frequency resolution on the perceived quality of vocoder speech. This section sets forth several conclusions concerning time-frequency resolution and speech quality. These conclusions were drawn from the results of the Category Judgement evaluation of the vocoded speech signals obtained from the simulation phase of the research.

It is hoped that these conclusions will contribute to the further understanding of the speech perception process and that they will point the way toward improvements in vocoder design. In addition, these conclusions should suggest areas for further research.

Conclusion One: Adequate Vocoder Time Resolution

Figure 3 shows the MCJ's for the nonadaptive vocoder plotted as a function of frame interval and cepstrum length. For each condition of frequency resolution, the two shorter frames showed an advantage in quality over the 40.0 ms frame. The performances of the 10.0 ms frame and the 20.0 ms frame were comparable for all cepstrum lengths except 4.0 ms where the 20.0 ms frame was signficantly better. This difference is unexplained but may be the results of some sort of time-frequency trading or simply a "quirk" in the data. It should be noted that the quality at a cepstrum length of 1.0 ms was essentially the same for all three frames, suggesting that poor frequency resolution was the predominant factor at that point.

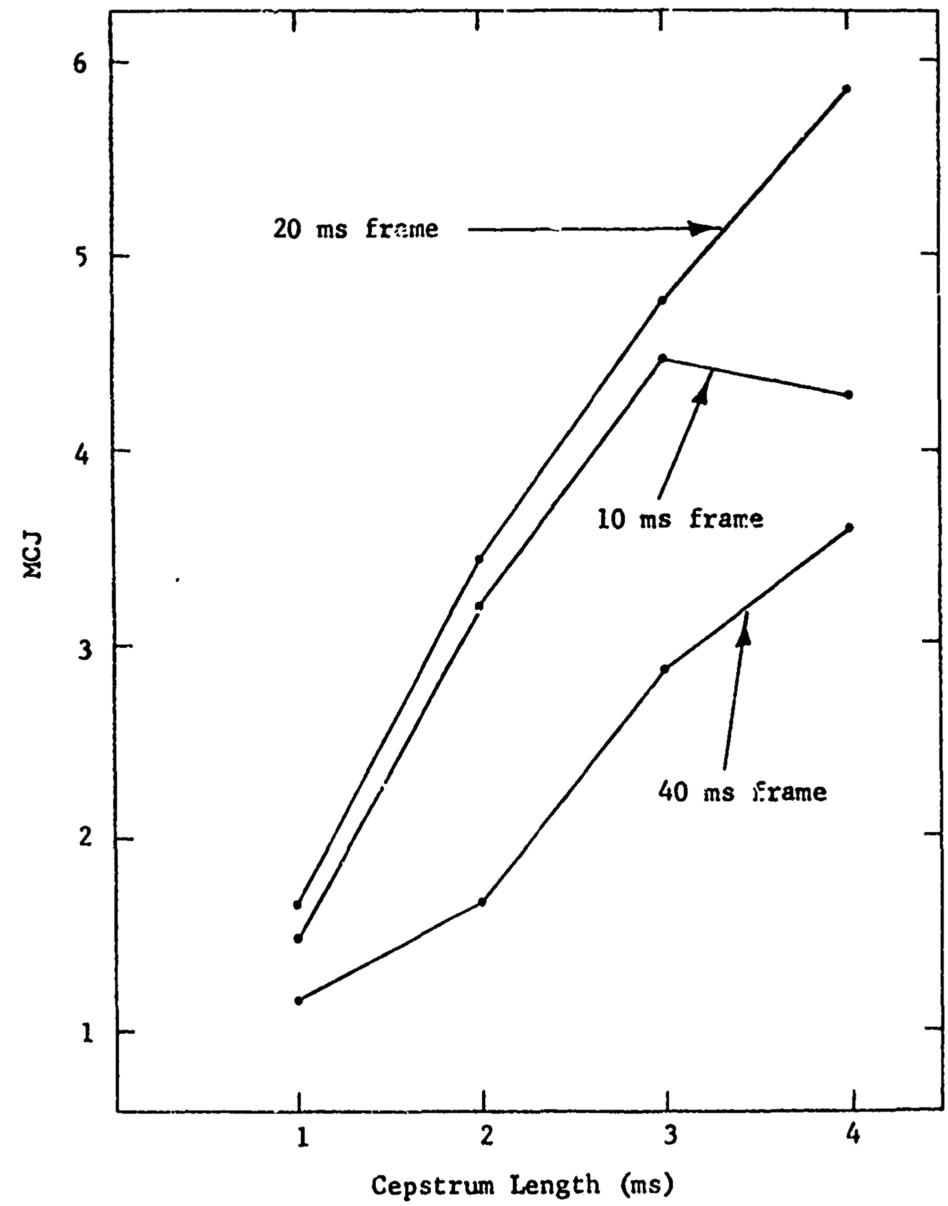

Figure 3. Performance of Nonadaptive Configurations

The conclusion to be drawn from this set of results is that maintaining time resolution better than about 20.0 ms seems to provide no improvement in speech quality.

Conclusion Two: Time-Frequency Trading in Quality Perception

An examination of the nonadaptive vocoders with 20.0 ms and 40.0 ms frames in Figure 3 shows that configurations with equivalent data rates (for example, 40.0 ms frame and 4.0 ms cepstrum compared to 20.0 ms frame and 2.0 ms cepstrum) have roughly equivalent quality. This observation may be interpreted as evidence of time-frequency trading in speech perception.

Conclusion Three: The Effect of Reduced Frequency Resolution in Unvoiced and Transition Regions

Figure 4 and 5 show plots of the MCJ's for the various adaptive vocoder configurations. Note that reducing the Mode 2 cepstrum length had no appreciable effect on the speech quality. This result was independent of Mode 1 frame, Mode 1 cepstrum, and Mode 2 frame. The conclusion is that frequency resolution can be reduced considerably in unvoiced regions and regions of voiced-unvoiced or unvoiced-voiced transition with little or no loss in speech quality. This is an important result for the design of adaptive vocoders which must maintain a constant data rate. This conclusion also lends support to the notion of time-frequency trading in speech perception.

Conclusion Four: The Effect of the Adaptive Strategy

The effect of the adaptive strategy is displayed in Figure 6. This figure gives a comparison of the MCJ's for several versions of the adaptive and nonadaptive vocoders. For the 20.0 ms frame at a fixed cepstrum

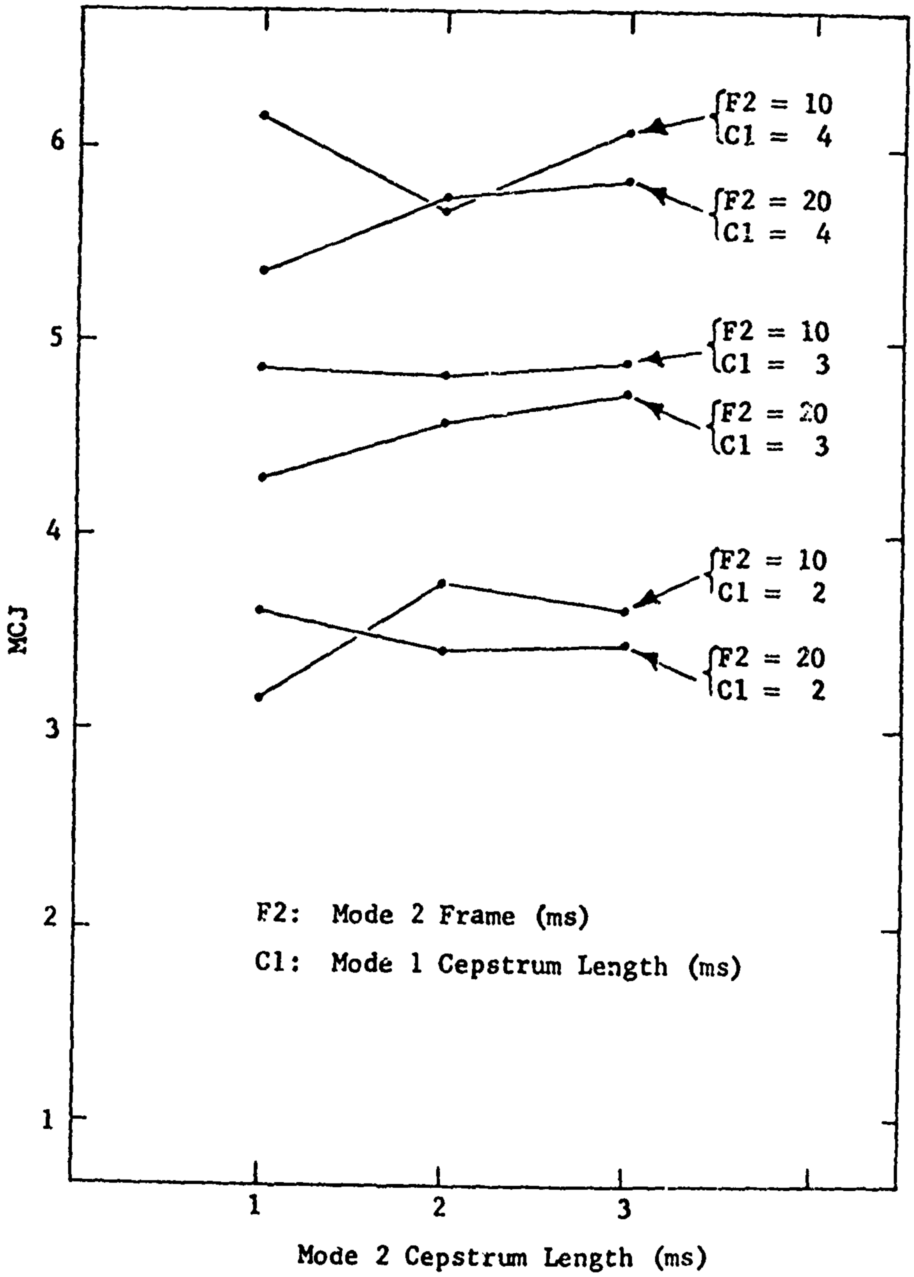

Figure 4. Performance of Adaptive Configurations with 20 ms Mode 1 Frame

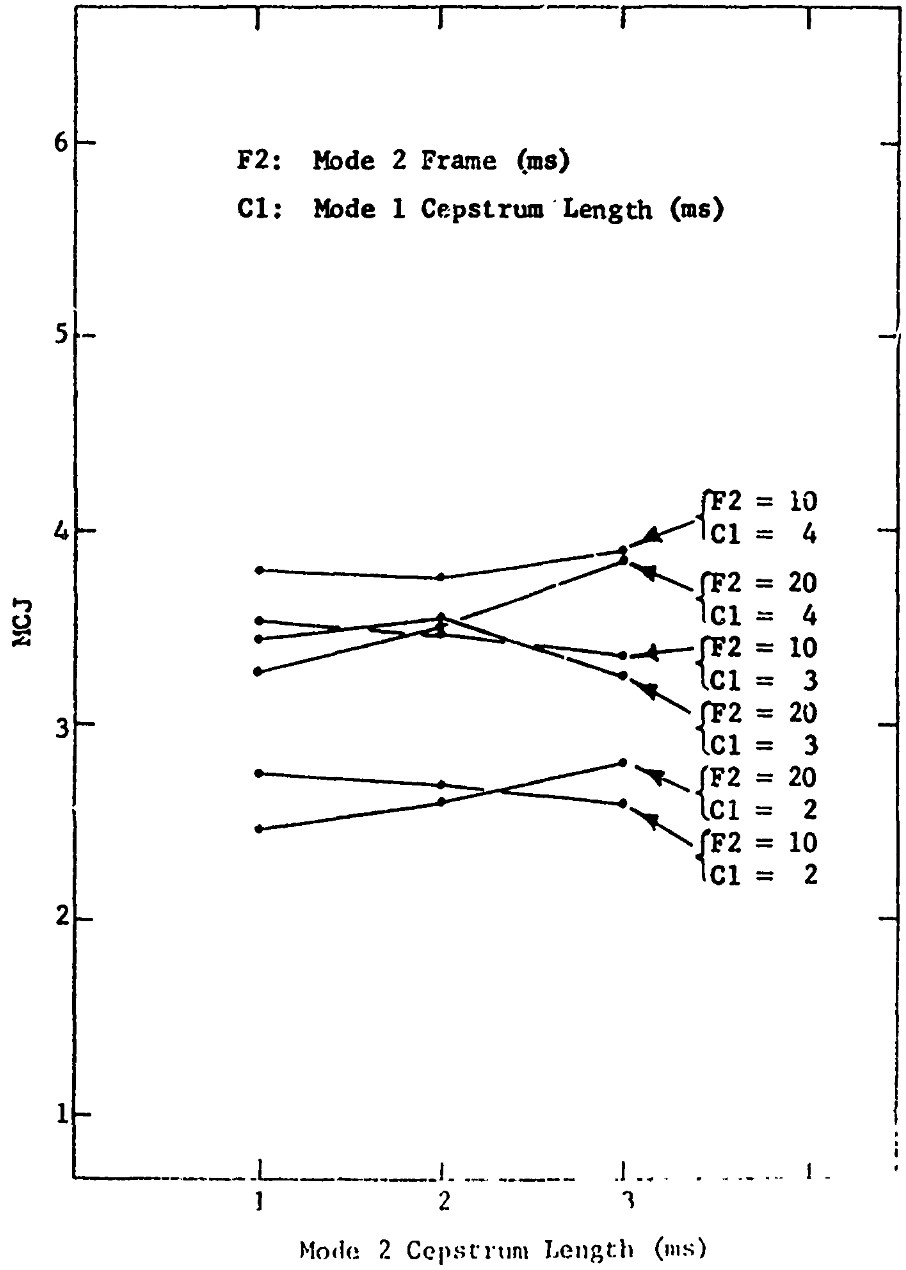

Figure 5. Performance of Adaptive Configurations
with 40 ms Mode 1 Frame

length, adapting to the 10.0 ms frame for unvoiced and transition regions resulted in no improvement in quality. This observation is in agreement with Conclusion One and suggests that time resolution of about 20.0 ms is sufficient for vocoders.

For the 40.0 ms frame and 4.0 ms cepstrum, adaption has no noticeable effect. However, for the 40.0 ms frame and either the 3.0 ms or 2.0 ms cepstrum, adapting to a shorter frame in unvoiced and transition regions led to improved quality. The improvement obtained was independent of the Mode 2 frame used, giving still more support to Conclusion One. The improvement was roughly equivalent to increasing the cepstrum length of the nonadaptive vocoder by 1.0 ms. It is not clear why adaption produced no improvement in quality for the 4.0 ms cepstrum case.

It appears that 20.0 ms time resolution is adequate for vocoder applications. Thus adaption in a system which normally maintains 20.0 ms resolution or better yields no improvement in performance. For systems that do not normally employ such good time resolution, adaption seems to offer considerable potential.

It should be pointed out that interpolation of impulse responses would probably improve the performance of the vocoders using 40.0 ms frames in voiced regions. Thus a combination of interpolation and adaption in these systems might well produce good quality speech at quite low data rates.

Discussion

The conclusions presented above should be regarded as tentative for several reasons. Only two speakers and four utterances were used in the research. Only one type of vocoder was employed and speech quality was

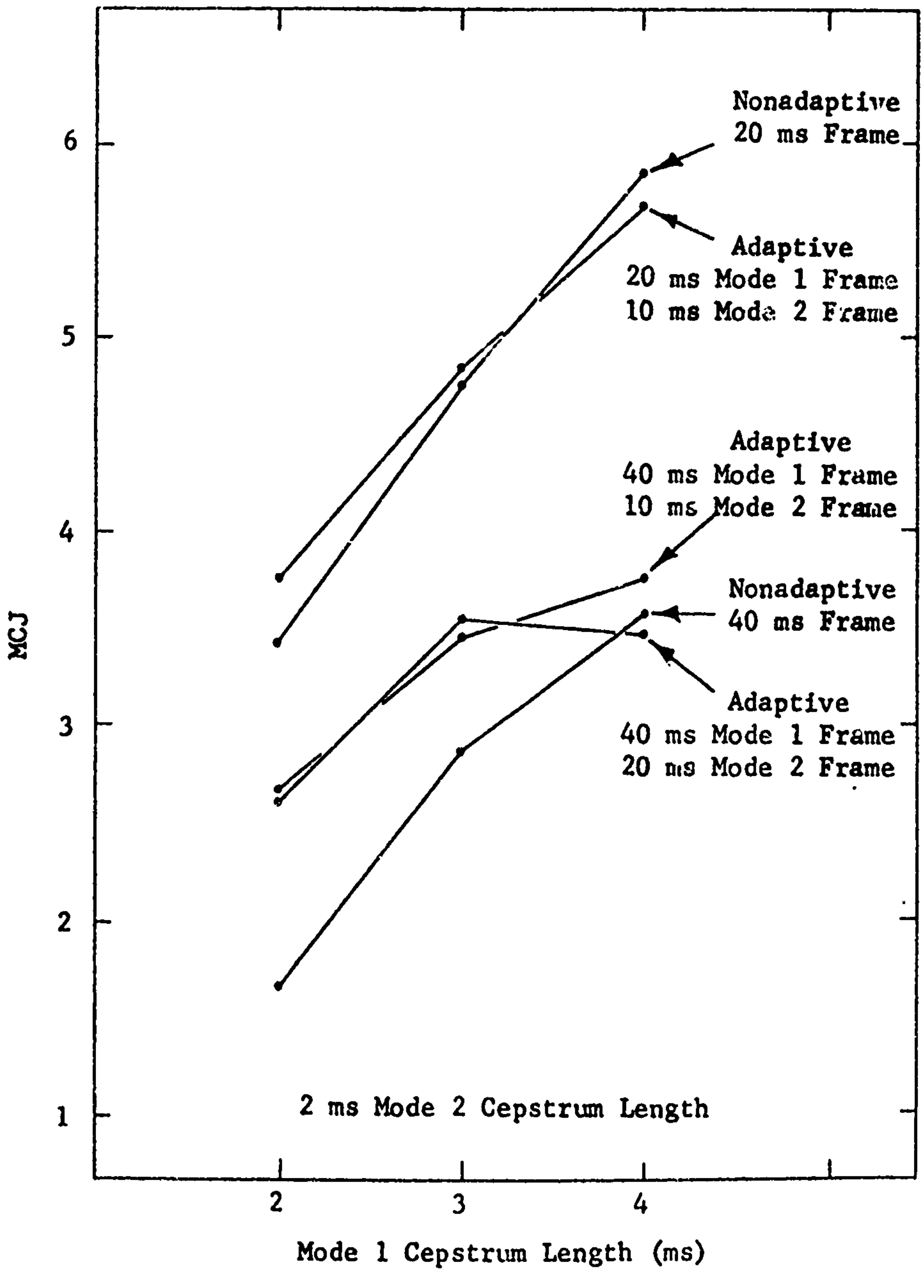

Figure 6. Effect on Adapting to a Shorter Frame in Unvoiced and Transition Regions

judged by only one of several methods available. It is certainly conceivable that a similar study with different source material, a different vocoder, or a different quality measurement technique could produce different results.

At an early stage in this research it became evident that the two speakers used produced vocoded speech of widely differing quality. The male speaker rated higher in quality than the female in all instances except three. This dependence of quality on the speakers may have colored the results of this study.

If a speaker sex dependent quality difference does exist in vocoder speech, it may be possible to explain it in terms of fundamental periods. Females typically have shorter fundamental periods than males so that there is more overlap of impulse responses during voiced speech and deconvolution is more difficult. An equivalent explanation can be given in the frequency domain. Since female speakers generally have shorter fundamental periods, the spectral lines in the short-time spectrum are spaced further apart so that the envelope due to the vocal tract is "sampled" less often in frequency making deconvolution by smoothing more difficult. In the particular case of the cepstrum vocoder, short fundamental periods cause the excitation portion of the cepstrum to encroach on the vocal tract portion with a resulting loss of quality in the deconvolution.

Influence of Pitch Signal Quality

Pitch, for the purposes of speech signal analysis, can be defined as the spacing between impulses in the excitation signal during voiced segments. This definition, though adequate for speech signals, ignores

effects observed by some in perception of pitch of higher frequency complex tones.[23]

The test results obtained in the study of time-frequency resolution[21] and summarized above utilize a "hand-painted" or perfect pitch excitation signal.

Hand-painting is done best by interactively studying the input speech signal using the dedicated interactive simulation facility.[24] A segment of speech is displayed on a CRT, together with displays of its spectrum, the output of a cepstrum pitch estimator, the output of an autocorrelation pitch estimator, and the output of a minimum difference pitch estimator. The operator then examines the time waveform and all the available pitch estimates and inputs his own choice of pitch period for that segment based on all available information.

The perceived speech quality of the synthesized speech, not only for the homomorphic vocoder but for any vocoder or LPC algorithm for speech synthesis, is found to be more sensitive to the quality of the pitch signal than to any other single parameter. Perturbations in the pitch signal, using known terms of pitch estimation, consist of small local offsets and large global errors or jumps. The large errors are often harmonically related as doubling, tripling, or halving of the pitch period. Local errors can cause a "wavering" quality or uncertainness in the synthetic speech. Global errors can cause total distortion including loss of intelligibility as well as reduced quality. Synthetic speech utilizing artificial pitch information, as in reading machines, produces a very unnatural or machine like quality.

PUBLICATIONS AND TECHNICAL REPORTS

During the period covered by this grant a number of publications and reports have been generated. These are listed below. Additional support for speech related work was obtained during the grant period. The list of reports includes the results of these studies as well.

To be submitted for publication:

"A Time-Frequency Resolution Experiment in Speech Analysis and Synthesis," C. R. Patisaul and J. C. Hammett.

Submitted for publication:

"The Multiband Pitch Detector," C. R. Patisaul and T. P. Barnwell.

Presented at Conferences and Published in Conference Record:

"Gapped ADPCM for Speech Digitization," T. P. Barnwell and A. M. Bush, NEC '74 Conference Record, October, 1974.

"A Minicomputer Based Digital Signal Processing System," T. P. Barnwell and A. M. Bush, EASCON '74 Conference Record, October, 1974.

Theses:

"Adaptive Time-Frequency Resolution in Vocal Tract Parameter Coding for Speech Analysis and Synthesis," C. R. Patisaul, Ph.D. Thesis, June, 1974.

Reports:

"Adaptive Differential PCM Speech Transmission," T. P. Barnwell, A. M. Bush, J. B. O'Neal, and R. W. Stroh, RADC-TR-74-177, Final Report, July, 1974.

"Pitch and Voicing in Speech Digitization," T. P. Barnwell, J. E. Brown, and C. R. Patisaul, Georgia Institute of Technology, School of Electrical Engineering, Research Report E21-620-74-BU-1, August, 1974.

"Recursive Algorithms for Data Processing," J. E. Brown, Georgia Institute of Technology, School of Electrical Engineering, Internal Memorandum Report, August, 1974.

PARTICIPATING PERSONNEL

The following personnel participated in the research programs conducted under this grant. The list is divided into faculty and students. Those students who earned an advanced degree during the course of the grant have listed with their name the degree and the date of the degree.

Faculty

T. P. Barnwell, Assistant Professor

A. M. Bush, Professor, Principal Investigator

Students

C. R. Patisaul, Ph.D., September, 1974

H. B. Brown, M.S.E.E., March, 1973

C. W. Stover, Ph.D. candidate

Involved to a lesser degree, and with no financial support from the grant, have been J. E. Brown, Assistant Professor, who participated as a consultant on theoretical issues, and several senior undergraduate project students who have participated in hardware projects in developing the dedicated minicomputer facility.

REFERENCES

1. Flanagan, J. L., Speech Analysis Synthesis and Perception, Academic Press, New York, 1965.

2. Fant, G., Acoustic Theory of Speech Production, Mouton and Co., The Hague, Netherlands, 1960.

3. Flanagan, J. L., Coker, C. H., Rabiner, L. R., Schafer, R. W., and Umeda, N., "Synthetic Voices for Computers," IEEE Spectrum, Vol. 7, No. 10, pp. 22-45, October, 1970.

4. Oppenheim, A. V., "Speech Analysis-Synthesis System Based on Homomorphic Filtering," J. Acoust. Soc. Amer., Vol. 45, No. 2, pp. 458-465, Februrary, 1969.

5. Oppenheim, A. V., "Speech Spectograms Using the Fast Fourier Transform," IEEE Spectrum, Vol. 7, No. 8, pp. 57-62, August, 1970.

6. Oppenheim, A. V., and Schafer, R. W., "Homomorphic Analysis of Speech," IEEE Trans. on Audio and Electroacoustics, Vol. AU-16, No. 2, pp. 221-226, June, 1968.

7. Weinstein, C. J., and Oppenheim, A. V., "Predictive Coding in a Homomorphic Vocoder," Preprint.

8. Gold, B., and Rader, G. M., Digital Processing of Signals, McGraw-Hill, New York, 1969.

9. Oppenheim, A. V. and Schafer, R. W., Digital Signal Processing, Prentice-Hall, 1975.

10. Makhoul, J. I. and Wolf, J. J., "Linear Prediction and the Spectral Analysis of Speech," Bolt, Beranek and Newman, Inc., Report BBN-2304, August, 1972.

11. Brown, J. E., "Recursive Algorithms for Data Processing," Georgia Institute of Technology, School of Electrical Engineering, Internal Memorandum Report, August, 1974.

12. Deczky, A. G., "Synthesis of Recursive Filters Using the Minimum p-Error Criterion," IEEE Trans. on Audio and Electroacoustics, Vol. AU-20, No. 4, pp. 257-264, October, 1972.

13. Kuo, F. F., and Kaiser, J. F., System Analysis by Digital Computer, John Wiley and Sons, Inc., 1966, pp. 218-244.

14. Boll, S. F., "A Priori Digital Speech Analysis," Advanced Research Projects Agency Report, AD762-029, and Ph.D. Thesis, University of Utah, March, 1973.

15. Melsa, J. L., et.al., "Development of a Configuration Concept of a Speech Digitizer Based on Adaptive Estimation Techniques," Final Report, Contract No. DCA 100-72-C-0036, Defense Communications Agency, August, 1973.

16. Atal, B. S. and Hanauer, S. L., "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," The Journal of the Acoustical Society of America, Vol. 50, No. 2, August 1971, pp. 637-665.

17. Atal, B. S. and Schroeder, M. R., "Adaptive Predictive Coding of Speech Signals," Bell System Technical Journal, Vol. 49, No. 8, October, 1970, pp. 1973-1987.

18. Markel, J. D. and Gray, A. H., Jr., "On Autocorrelation Equations as Applied to Speech Analysis," IEEE Trans. on Audio and Electroacoustics, Vol. AU-21, No. 2, April 1973, pp. 69-80.

19. Itakura, F. and Saito, S., "On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer," Conference Record, IEEE 1972 Conference on Speech Communication and Processing, IEEE Catalog No. 72-CHO-596-7-AE, April 1972, pp. 434-437.

20. Dunn, J. G., "An Experimental 9600-Rit/s Voice Digitizer Employing Adaptive Prediction," IEEE Trans. on Communication Technology, Vol. COM-19, No. 6, December 1971, pp. 1021-1033.

21. Patisaul, C. R., "Adaptive Time-Frequency Resolution in Vocal Tract Parameter Coding for Speech Analysis and Synthesis," Ph.D. Thesis, Georgia Institute of Technology, June, 1974.

22. Barnwell, T. P. and Bush, A. M., "A Minicomputer Based Digital Signal Processing System," EASCON '74 Conference Record, October 1974.

23. Wightman, F. L. and Green, D. M., "The Perception of Pitch," American Scientist, Vol. 62, No. 2, March-April, 1974, pp. 208-215.

24. Barnwell, T. P., Brown, J. E., Bush, A. M., and Patisaul, C. R., "Pitch and Voicing in Speech Digitization," Georgia Institute of Technology, School of Electrical Engineering, Research Report E-21-620-74-BU-1, August, 1974.